Компьютерная лингвистика и интеллектуальные технологии:
по материалам международной конференции «Диалог 2017»

Москва, 31 мая — 3 июня 2017

КАРТА СЛОВ: ПЕРЕОСМЫСЛЕНИЕ ПОДХОДА К СОСТАВЛЕНИЮ ОНЛАЙН-СЛОВАРЕЙ В ПОСТМОБИЛЬНУЮ ЭРУ

Кулагин Д. И. (kdenisk@gmail.com)

Карта слов (https://kartaslov.ru/), Москва, Россия

В статье мы хотим рассказать об эволюции подходов к поиску словарной информации от эпохи бумажных словарей до настоящего времени, когда большая часть информации в интернете просматривается с персональных мобильных устройств. В тексте освещаются современные походы к структурированию словарных записей, даётся обзор существующих решений, приводится анализ их сильных и слабых сторон.

Во второй части статьи рассказывается о созданном автором интернет-ресурсе «Карта слов», разработанных в рамках проекта лингвистических технологиях, а также новых подходах к решению проблем, описанных в первой части статьи.

Решаемая проблематика

В эпоху бумажных словарей составителям приходилось решать проблему ограниченного количества печатного места, т. к. это напрямую влияло на стоимость издания и удобство использования. В результате была придумана система помет, сокращений и перекрёстных ссылок, позволявшая экономить печатный ресурс.

Одним из ярких примеров детально проработанной системы кодирования словарной информации является грамматический словарь русского языка Зализняка А. А. Первые 142 страницы издания представляют собой подробнейшую инструкцию по пользованию словарём. О сложностях такого подхода для рядового читателя поговорим ниже.

Развитие вычислительной техники и распространение персональных компьютеров создало естественную потребность к оцифровке бумажных словарей, что и было сделано. В процессе внесения информации в компьютер вся система обозначений и навигации была скопирована в машину практически без изменений.

В результате мы получили тот же бумажный словарь, но в цифровом виде.

Изменение подхода к поиску и обнаружению информации

В доцифровую эпоху единственным способом найти словарную статью было достать с полки бумажное издание и, используя алфавитный индекс, выйти на соответствующую страницу.

С появлением персональных компьютеров распространение получили автономные программы, устанавливаемые на ПК пользователя и позволявшие осуществлять быстрый поиск по индексу.

Ситуация кардинально поменялась с распространением доступного широкополосного Интернета, когда необходимость в установке отдельного программного обеспечения фактически отпала — появилась возможность поиска словарных статей в онлайн-пространстве.

Ещё одной важной точкой в изменении подходов к поиску информации стало широкое распространение персональных мобильных устройств. По статистике нашего проекта, доля мобильных пользователей составляет чуть более 50 % и продолжает неуклонно увеличиваться.

Итак, сегодня основной точкой входа в словарную статью является переход из поиска. В большинстве случаев это поисковые системы, такие как Яндекс и Гугл. В меньшей степени — собственный поиск на сайте.

Т.о. любой словарный проект, который рассчитывает быть заметным в современном информационном пространстве, должен быть встроен в описанную выше систему поиска и обнаружения информации. Не в последнюю очередь иметь структуру, дружественную поисковым системам.

Обновлённые требования к системе обозначений

Все бумажные словари подразумевают кривую обучения пользования изданием. Как минимум человеку придётся ознакомиться с системой помет и способом упорядочивания словарных статей.

В большинстве случаев такая система довольно стандартна и освоив её для одного словаря, человек может использовать полученные знания при работе с другими изданиями.

Для онлайн-словаря потребность в системе помет, а также их разумном сокращении сохраняется. Тем не менее разработчики имеют возможность добавить интерактивные подсказки и отказаться от использования сложных для восприятия сокращений. Теперь основная цель — удобство читателя и уменьшение до минимума длины кривой обучения.

Словоизменительные парадигмы, к примеру, можно публиковать полностью, в развёрнутом виде, минуя необходимость для рядового пользователя вникать в сложную систему кодирования.

Изменение подхода к специализации

По причинам, описанным выше, бумажные словари чаще всего имели чёткую специализацию. Орфографические, орфоэпические, толковые, словари синонимов и т. д.

В цифровом пространстве необходимость экономии печатного места естественным образом отпадает. Зато появляется новая — каким образом структурировать и приоритизировать возросший объём информации, чтобы пользователю было удобно.

Переосмысление подхода к структуре словарной статьи на примере Викисловаря

Удачным, на взгляд автора, примером переосмысления подхода к структуре словаря и словарной статьи является Викисловарь. Он сочетает в себе тесную интеграцию с поисковыми системами и широкую специализацию представленных словарных записей.

Для слов и выражений, представленных в Викисловаре доступна следующая информация:

- ударение и парадигма склонения/спряжения;
- морфологическая информация;
- разбор по составу;
- толкования с примерами употребления в литературе;
- синонимы и антонимы;
- гиперонимы и гипонимы;
- родственные слова;
- этимология;
- фразеологизмы и устойчивые сочетания.

Словарные записи в Викисловаре становятся универсальными, а сам словарь включает в себя семантические отношения, т. е. одновременно выполняет роль тезауруса.

Слабые стороны Викисловаря

1. Викисловарь открыт для редактирования каждому человеку и пополняется сообществом. Соответственно представленная информация не является авторитетной и в определённой доле случаев оказывается неточной/неверной.
2. В основе Викисловаря, как технологии, лежит универсальный текстовый викидвижок. Соответственно целостность структуры словаря и информации в нём поддерживается только на основании договорённостей, принятых в рамках сообщества. За исключением редких случае программный контроль не осуществляется.
3. Викисловарь не использует компьютерные технологии автоматического извлечения лингвистической информации, что могло бы сильно облегчить труд составителей.
4. Словарь не является «живым» в том смысле, что он автоматически не перестраивается по мере использования.

Общая проблематика онлайн-словарей в российском сегменте Интернета

С изменением подходов к поиску и обнаружению информации, возникла острая потребность в новых цифровых словарях, доступных онлайн. Спрос родил предложение и в настоящее время в Интернете можно обнаружить массу сайтов, предлагающих лингвистическую информацию различного качества.

Если не брать в расчёт крупные проекты, наподобие Викисловаря, то в основной массе такие ресурсы представляют собой оцифрованные копии распространённых словарей Шведовой Н. Ю., Ефремовой Т. Ф., Ожегова С. И. и других известных лингвистов, компиляцию из различных источников, а также сгенерированную компьютером информацию.

Существующие ресурсы в большой степени закрывают потребность в толковании слов, но такие вопросы как правильное ударение, допустимость нескольких вариантов постановки ударения и постановка ударения в косвенных падежах имени существительного, а также финитных форм глагола, освещены в них недостаточно полно и/или достоверно.

Карта слов: переосмысление подхода к составлению онлайн-словарей в постмобильную эру

В предыдущих разделах настоящей статьи мы постарались очертить проблематику современных цифровых словарей, указали на существующие решения, проанализировали их сильные и слабые стороны.

При создании Карты слов мы поставили задачу создать онлайн-словарь и интегрировать его в современное информационное пространство.

Структура словаря и система навигации

В основу навигации по сайту положены так называемые «карты слов». По сути карта отдельного слова или выражения представляет собой краткую выжимку всех сведений о лексической единице — подобие словарной записи в Викисловаре.

Карта содержит ссылки на более подробную информацию по отдельным срезам словарной статьи:

- толкование;
- ассоциации;
- синонимы;
- примеры употребления в контексте;
- контекстные связи;
- справочная информация по склонению/спряжению;
- справочная информация по морфемному строению.

Все отдельные подразделы содержат перекрёстные ссылки друг на друга, обеспечивая прозрачную навигацию между ними.

Выбор такой структуры словаря обусловлен внимательным анализом поведения пользователей. Как правило посетитель, изначально искавший

примеры употребления в контексте, и дальше остаётся в том же подразделе. Это позволяет словарю быть универсальным, но при этом сохранить удобство пользования, как если бы он имел чёткую специализацию.

Больш*ая* часть внутренних переходов осуществляется с использованием поиска. Для удобства пользователей поиск снабжён системой подсказок, учитывает морфологию русского языка и корректирует фонетические ошибки при написании слов.

Востребованность собственной системы поиска на сайте достаточно велика — ей пользуется как минимум один из пяти посетителей сайта.

Лингвистический движок и современные технологии компьютерной обработки языка

В основе сайта лежит собственный лингвистический движок, разработанный специально для Карты слов. Это позволяет обеспечивать высокую точность информации, представленной на сайте, а также гибкую навигацию между словарными статьями.

Лингвистический движок используется для следующих целей:

1. Верификация представленной на сайте информации, автоматический поиск орфографических ошибок и опечаток, фильтрация нежелательного контента на всех уровнях.
2. Проверка соответствия справочной информации правилам русского языка, поиск аномалий для осуществления дополнительного контроля человеком.
3. Поиск примеров употребления слов и выражений в контексте с учётом морфологии. Текущая реализации работает со скоростью около 2500 поисков/секунду.
4. Автоматическая разбивка «мешка синонимов» на синсеты.

Важнейшее преимущество, которое даёт использование лингвистического движка — постепенное наращивание понимания машиной структуры языка, повышение достоверности справочной информации и возможность компьютерной поддержки составителей и редакторов словаря.

Потенциал для использования технологий компьютерной обработки языка

При работе над Картой слов мы выявили ряд задач, которые нецелесообразно или сложно решать вручную:

- составление словаря синонимов для узкоспециализированных тематик;
- составление словаря синонимов и сходных по смыслу выражений для словосочетаний;
- выделение контекстных связей.

Все вышеперечисленные задачи можно с определённой точностью решить автоматически, создавая при этом для пользователя дополнительную ценность.

Открытые данные

Часть структурированных данных, полученных в ходе работы над словарём, может оказаться полезной специалистам и исследователям-лингвистам. Отличным примером использования открытых данных для развития технологий обработки русского языка является библиотека *pymorphy2,* в работе которой используется словарь *OpenCorpora*.

В рамках работы над Картой слов мы опубликовали набор данных, включающий сделанные реальными людьми орфографические ошибки и опечатки. В ближайшем будущем планируется также опубликовать словарь ассоциаций и используемый морфологический словарь, если он окажется более полным, чем уже имеющиеся в открытом доступе.

Планы на будущее

В рамках Карты слов был разработан и успешно опробован ряд лингвистических технологий, найдена положительно воспринятая пользователями структура словаря и система навигации.

В настоящий момент мы ищем возможность наладить взаимодействие с профессиональными лингвистами, лексикографами и составителями словарей с тем, чтобы создать в русскоязычном сегменте Интернета авторитетный словарь русского языка, опирающийся на традиционно сильную советско-российскую лингвистическую школу и современные компьютерные и медиа подходы.

Со своей стороны, мы готовы предоставить ключевые технологии, понимание устройства современного интернет-пространства, заинтересованную аудиторию.

Контакты

1. Сайт Карты слов — https://kartaslov.ru/
2. Открытые данные — https://github.com/dkulagin/kartaslov
3. Идея и реализация проекта, вопросы сотрудничества — Кулагин Денис — kdenisk@gmail.com